Why dialects differ: The influence of concept features on lexical geographical variation

Karlien Franco, Dirk Speelman & Dirk Geeraerts

RU Quantitative Lexicology and Variational Linguistics

Lexical variation

semasiology
meaning

cookie

onomasiology
naming

present

gift

Lexical variation

semasiology
meaning

cookie

onomasiology
naming

present

gift

Onomasiological heterogeneity

= synonymy across dialects

→ onomasiological perspective:

why do certain concepts

show more variation than others?

Pilot study

Geeraerts and Speelman (2010)
Speelman and Geeraerts (2008)

(heterodox) concept features influence onomas. heterogeneity

- more geographical variation for non-salient concepts

e.g. SLUIK HAAR (‘straight hair’) is less salient than HOOFD (‘head’)

- more geographical variation for vaguer concepts

e.g. LIES (‘groin’) is more vague than DUIM (‘thumb’)

- more geographical variation for negatively connoted concepts

e.g. KWIJL (‘drool’) vs. JUKBEEN (‘cheekbone’)

Pilot study

Geeraerts and Speelman (2010)

Speelman and Geeraerts (2008)

(heterodox) concept features influence onomas. heterogeneity

- variation reflects geographical patterns
- also reflects variability that is inherent to semantic features of concepts

→ can also influence dialectometric analyses

Pilot study

Geeraerts and Speelman (2010)

Speelman and Geeraerts (2008)

data

- the human body
- digitized database of Dictionary of Limburgish dialects

methods

- linear regression
- dialectometric analysis

Aim

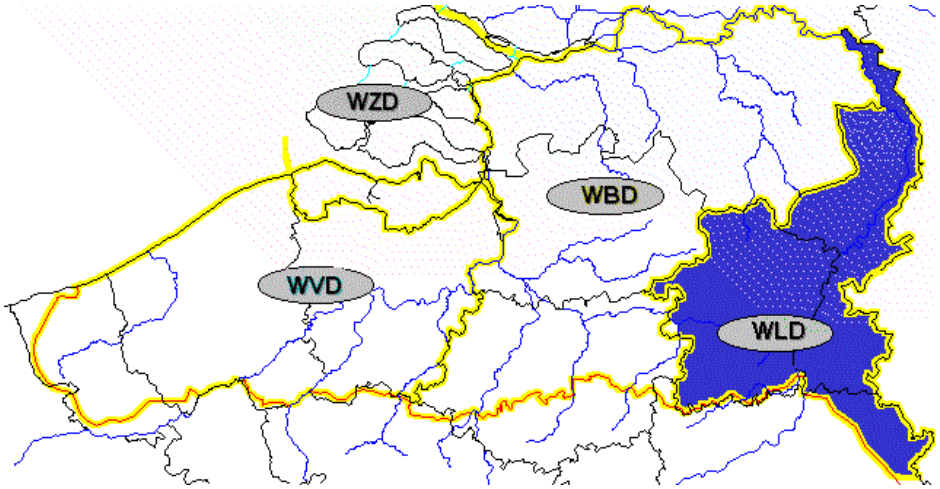
- replicate pilot study
- expand scope to other semantic fields

→ additional RQ:

influence of semantic field on onomasiological heterogeneity?

Data

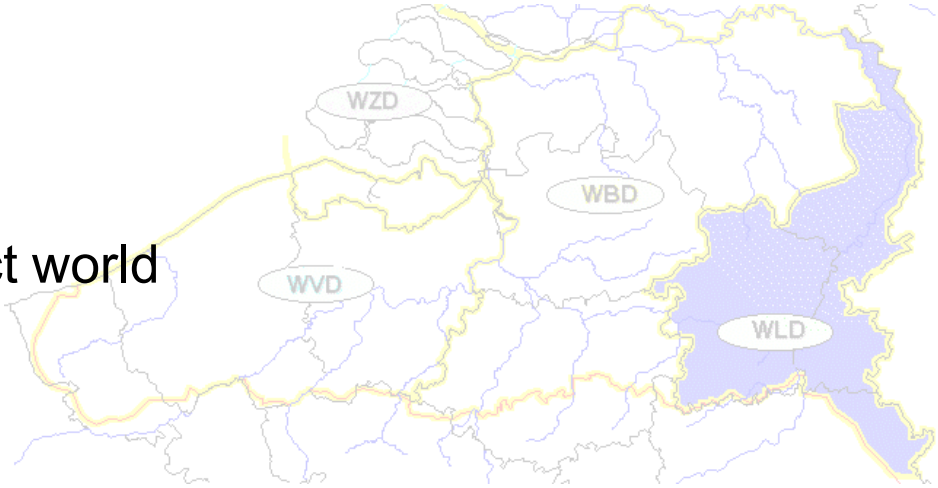
- digitized database regional Limburgish dialect dictionary (‘Woordenboek van de Limburgse dialecten’)



Data

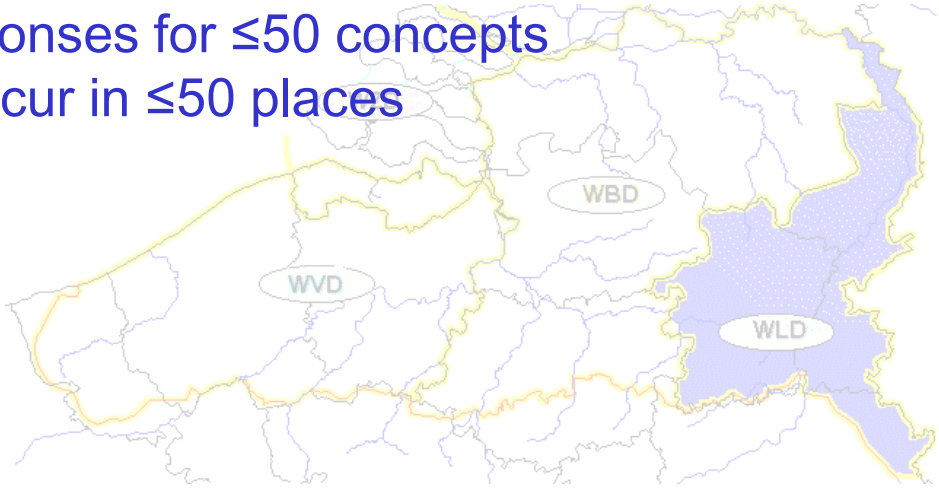
- digitized database regional Limburgish dialect dictionary (‘Woordenboek van de Limburgse dialecten’)
  - consists of 3 parts:
    1. agrarian terminology (13 semantic fields)
    2. non-agrarian professional terminology (12 sem. fields)
    3. general vocabulary (14 semantic fields)

  - use 4 semantic fields from general vocab:
    - the human body
    - character and feelings
    - family & sexuality
    - the physical and abstract world



Data

- digitized database regional Limburgish dialect dictionary ('Woordenboek van de Limburgse dialecten')
  - data from Belgium & the Netherlands
  - based on questionnaire data and on smaller local dictionaries
    → focus on consistent questionnaire data (N*)
  - noisy data
    → exclude places with responses for ≤50 concepts
    and concepts that occur in ≤50 places



Overview

Introduction

Defining concept features

Defining onomasiological heterogeneity

Results

Conclusion

Overview

Introduction

Defining concept features

Defining onomasiological heterogeneity

Results

Conclusion

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

Results

Conclusion

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

Results

Conclusion

1. Semantic field of the concept

- 4 parts of digitized dictionary of Limburgish dialects (general vocabulary):
  - character and feelings

  e.g. *bangerik* ('coward'), *plezier maken* ('to have fun')
  - family and sexuality

  e.g. *huwelijk* ('wedding'), *luier* ('nappy', 'diaper')
  - the human body

  e.g. *oorlel* ('ear lobe'), *gekruld haar* ('curly hair')
  - the physical and abstract world

  e.g. *millimeter*, *voet* ('foot'), *motregenen* ('to drizzle')

- we expect little variation in semantic field *the physical and abstract world*: objective concepts with few connotations show less variation

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

Results

Conclusion

2. Lack of salience of the concept

~ how well known is the concept?

we expect less salient concepts to show more onomas. heterogeneity

best operationalization of (lack of) salience?

- relative number of multi-word expressions as responses per concept
- relative number of hesitant responses per concept
- number of places without a response per concept
- lack of prevalence per concept

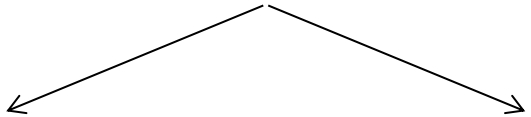
2. Lack of salience of the concept

- relative number of multi-word expressions as responses per concept

- relative number of hesitant responses per concept
- number of places without a response per concept
- lack of prevalence per concept

2. Lack of salience of the concept

- relative number of multi-word expressions as responses per concept



basic-level hypothesis: well-known concepts get shorter responses

hesitant MW responses to hide unfamiliarity with concept or with dialectal word for the concept

- relative number of hesitant responses per concept
- number of places without a response per concept
- lack of prevalence per concept

2. Lack of salience of the concept

- relative number of multi-word expressions as responses per concept
- relative number of hesitant responses per concept

1. coding procedure:
    - hesitant: hesitant responses, unclear cases
    - not hesitant: fixed expressions and all other responses
2. per concept proportion of hesitant responses (type level)

- number of places without a response per concept
- lack of prevalence per concept

2. Lack of salience of the concept

- relative number of multi-word expressions as responses per concept
- relative number of hesitant responses per concept
- number of places without a response per concept

  - number of places that occur in each part of the dictionary
  - number of places with a response for each concept

  → number of missing places per concept per dictionary part

- lack of prevalence per concept

2. Lack of salience of the concept

- relative number of multi-word expressions as responses per concept
- relative number of hesitant responses per concept
- number of places without a response per concept
- lack of prevalence per concept

  - data from large-scale lexical decision experiment (Keuleers et al. 2015): “the proportion of a population knowing a particular word”
  - calculation:
    - match standard description of concepts to words in prevalence data
    - 1 – minimum of scores for Belgium and the Netherlands

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

Results

Conclusion

3. Concept vagueness

- we expect vague, overlapping concepts with fuzzy boundaries to show more onomasiological heterogeneity

- operationalized in terms of lexical non-uniqueness (cf. pilot study): per concept number of lexical types that occur for other concepts as well

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

Results

Conclusion

4. Affect

- we expect concepts with (negative?) affect to show more onomasiological lexical variation
- operationalization:
  - concept polarity
  - relative number of diminutive responses (type-level) per concept

4. Affect

- we expect concepts with (negative?) affect to show more onomasiological lexical variation
- operationalization:
  - concept polarity

    → coding procedure: sensitivity of each concept to negative / positive affect or lack thereof

  - relative number of diminutive responses (type-level) per concept

4. Affect

- we expect concepts with (negative?) affect to show more onomasiological lexical variation
- operationalization:
  - concept polarity
  - relative number of diminutive responses (type-level) per concept

  → coding procedure: automatic & checked manually

  → proportion of diminutive responses per concept (type level)

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

Results

Conclusion

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

1. diversity
2. geographic fragmentation
3. onomasiological heterogeneity

Results

Conclusion

1. Diversity

= number of variants per concept (type-level)

e.g. BEGRAFENIS (‘funeral’):

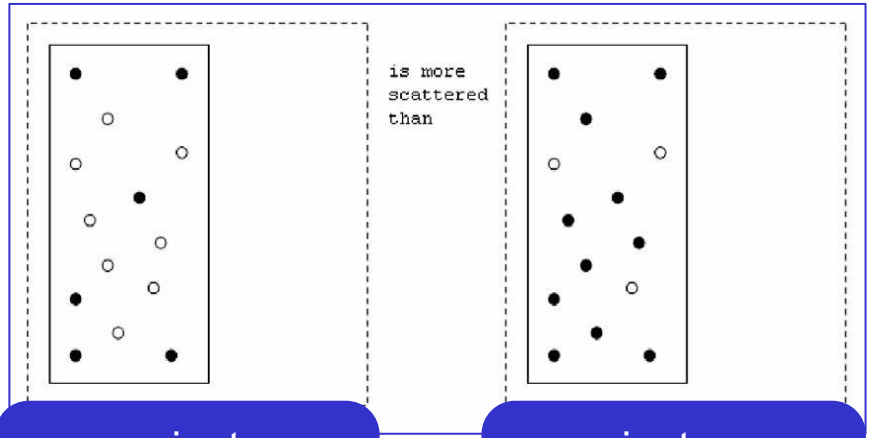
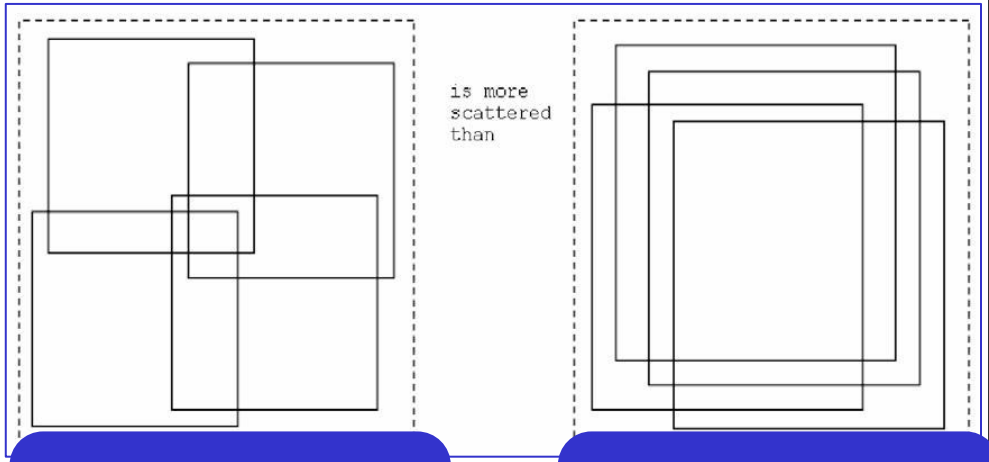
- beerdigung (du.)
- begrafenis
- begrbnis (du.)
- uitvaart

→ diversity = 4

2. Geographic fragmentation

~ lack of homogeneity in geographical spread of a concept

- consists of two parts:

dispersion	range
	
variants scattered across dialect area	each word type occurs in small geographical area
variants are found in nearby locations	each word type takes up almost entire dialect area

3. Onomasiological heterogeneity

- operationalization:

log(diversity * geographical fragmentation) per concept

- response variable for multivariate linear regression

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

1. diversity
2. geographic fragmentation
3. onomasiological heterogeneity

Methodology & results

Conclusion

Methodology

- data:
  - N = 145 951
  - total number of concepts = 859
  - total number of places = 244

- multivariate linear regression
  - response: onomasiological heterogeneity
  - predictors: operationalizations of semantic field, lack of salience, vagueness, affect

Results

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            3.36974    0.11559  29.152  < 2e-16 ***
SEMFIELDcharacter_and_feelings         0.05479    0.13933   0.393  0.69427
SEMFIELDfamily_and_sexuality          -0.76615    0.17356  -4.414 1.20e-05 ***
SEMFIELDthe_human_body                -0.84314    0. 14966 -5.634 2.68e-08 ***
LACKOFSALIENCE.relative.nr.mwe         2.80377    0.28364   9.885  < 2e-16 ***
LACKOFSALIENCE.lack.of.prevalence      1.71676    0.52333   3.280  0.00109 **
AFFECT.polaritynegative                1.02845    0.12753   8.064 3.84e-15 ***
AFFECT.polaritypositive                0.99574    0.22965   4.336 1.70e-05 ***
AFFECT.rel.nr.dimin                    1.49635    0.47909   3.123  0.00187 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 618 degrees of freedom
  (230 observations deleted due to missingness)
Multiple R-squared:  0.3273,    Adjusted R-squared:  0.3186
F-statistic: 37.59 on 8 and 618 DF,  p-value: < 2.2e-16
```

Overview predictors

RQ	hypothesis	operationalization
semantic field	less variation in *phys. and abstract world* than in other semantic fields	4 parts of dictionary
lack of salience	less salient concepts show more variation	- relative nr. MWE's/concept - relative nr. hesitant resp./conc. - nr of places without resp./conc. - lack of prevalence/conc.
vagueness	vaguer concepts show more variation	nr. of non unique types/concept
affect	(negatively?) connoted concepts show more variation	- concept polarity - nr. of diminutive responses/conc.

Overview predictors

RQ	hypothesis	operationalization
semantic field	less variation in *phys. and abstract world* than in other semantic fields	4 parts of dictionary
lack of salience	less salient concepts show more variation	- relative nr. MWE's/concept - relative nr. hesitant resp./conc. - nr of places without resp./conc. - lack of prevalence/conc.
vagueness	vaguer concepts show more variation	nr. of non unique types/concept
affect	(negatively?) connoted concepts show more variation	- concept polarity - nr. of diminutive responses/conc.

Results

```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          3.36974    0.11559  29.152  < 2e-16 ***
SEMFIELDcharacter_and_feelings       0.05479    0.13933   0.393  0.69427
SEMFIELDfamily_and_sexuality        -0.76615    0.17356  -4.414 1.20e-05 ***
SEMFIELDthe_human_body              -0.84314    0. 14966 -5.634 2.68e-08 ***
LACKOFSALIENCE.relative.nr.mwe       2.80377    0.28364   9.885  < 2e-16 ***
LACKOFSALIENCE.lack.of.prevalence    1.71676    0.52333   3.280  0.00109 **
AFFECT.polaritynegative
AFFECT.polaritypositive
AFFECT.rel.nr.dimin
---
Signif. codes:  0 '***' 0.001 '*

Residual standard error: 1.236 o
  (230 observations deleted due
Multiple R-squared:  0.3273,
F-statistic: 37.59 on 8 and 618
```

original hypothesis:
less variation in *phys. and abstract world* than in other semantic fields

results:
more onomasiological heterogeneity in semantic field *character and feelings*

Overview predictors

RQ	hypothesis	operationalization
semantic field	less variation in *phys. and abstract world* than in other semantic fields	4 parts of dictionary
lack of salience	less salient concepts show more variation	- relative nr. MWE's/concept - relative nr. hesitant resp./conc. - nr of places without resp./conc. - lack of prevalence/conc.
vagueness	vaguer concepts show more variation	nr. of non unique types/concept
affect	(negatively?) connoted concepts show more variation	- concept polarity - nr. of diminutive responses/conc.

Results

```
Coefficients:

                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          3.36974    0.11559  29.152  < 2e-16 ***
SEMFIELDcharacter_and_feelings       0.05479    0.13933   0.393  0.69427
SEMFIELDfamily_and_sexuality        -0.76615    0.17356  -4.414 1.20e-05 ***
SEMFIELDthe_human_body              -0.84314    0. 14966 -5.634 2.68e-08 ***
LACKOFSALIENCE.relative.nr.mwe       2.80377    0.28364   9.885  < 2e-16 ***
LACKOFSALIENCE.lack.of.prevalence    1.71676    0.52333   3.280  0.00109 **
AFFECT.polaritynegative              1.02845    0.12753   8.064 3.84e-15 ***
AFFECT.polaritypositive
AFFECT.rel.nr.dimin
---
Signif. codes:  0 '***' 0.001 '*

Residual standard error: 1.236 o
  (230 observations deleted due
Multiple R-squared:  0.3273,
F-statistic: 37.59 on 8 and 618
```

original hypothesis:
less salient concepts show more variation

confirmed by model:

- more variation for concepts with a higher proportion of MWE's
- more variation for concepts with a low prevalence score

Overview predictors

RQ	hypothesis	operationalization
semantic field	less variation in *phys. and abstract world* than in other semantic fields	4 parts of dictionary
lack of salience	less salient concepts show more variation	- relative nr. MWE's/concept - relative nr. hesitant resp./conc. - nr of places without resp./conc. - lack of prevalence/conc.
vagueness	vaguer concepts show more variation	nr. of non unique types/concept
affect	(negatively?) connoted concepts show more variation	- concept polarity - nr. of diminutive responses/conc.

Results

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            3.36974    0.11559  29.152  < 2e-16 ***
SEMFIELDcharacter_and_feelings         0.05479    0.13933   0.393  0.69427
SEMFIELDfamily_and_sexuality          -0.76615    0.17356  -4.414 1.20e-05 ***
SEMFIELDthe_human_body                -0.84314    0. 14966 -5.634 2.68e-08 ***
LACKOFSALIENCE.relative.nr.mwe         2.80377    0.28364   9.885  < 2e-16 ***
LACKOFSALIENCE.lack.of.prevalence      1.71676    0.52333   3.280  0.00109 **
AFFECT.polaritynegative                1.02845    0.12753   8.064 3.84e-15 ***
AFFECT.polaritypositive                0.99574    0.22965   4.336 1.70e-05 ***
AFFECT.rel.nr.dimin                    1.49635    0.47909   3.123  0.00187 **
---
Signif. codes:  0 '***'

Residual standard error
  (230 observations del
Multiple R-squared:  0.
F-statistic: 37.59 on 8
```

hypothesis:
vaguer concepts show more variation

results:
predictor VAGUENESS does not reach significance in this model

Overview predictors

RQ	hypothesis	operationalization
semantic field	less variation in *phys. and abstract world* than in other semantic fields	4 parts of dictionary
lack of salience	less salient concepts show more variation	- relative nr. MWE's/concept - relative nr. hesitant resp./conc. - nr of places without resp./conc. - lack of prevalence/conc.
vagueness	vaguer concepts show more variation	nr. of non unique types/concept
affect	(negatively?) connoted concepts show more variation	- concept polarity - nr. of diminutive responses/conc.

Results

```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          3.36974    0.11559  29.152  < 2e-16 ***
SEMFIELDcharacter_and_feelings       0.05479    0.13933   0.393  0.69427
SEMFIELDfamily_and_sexuality        -0.76615    0.17356  -4.414 1.20e-05 ***
SEMFIELDthe_human_body              -0.84314    0. 14966 -5.634 2.68e-08 ***
LACKOFSALIENCE.relative.nr.mwe       2.80377    0.28364   9.885  < 2e-16 ***
LACKOFSALIENCE.lack.of.prevalence    1.71676    0.52333   3.280  0.00109 **
AFFECT.polaritynegative              1.02845    0.12753   8.064 3.84e-15 ***
AFFECT.polaritypositive              0.99574    0.22965   4.336 1.70e-05 ***
AFFECT.rel.nr.dimin                  1.49635    0.47909   3.123  0.00187 **
---
Signif. codes:  0 '***'

Residual standard error
  (230 observations del
Multiple R-squared:  0.
F-statistic: 37.59 on 8
```

original hypothesis:
(negatively?) connoted concepts show more variation

results:

- negative AND positive concepts show more variation
- proportion of diminutives also reflects affect

Results

```
Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              3.36974    0.11559  29.152  < 2e-16 ***
SEMFIELDcharacter_and_feelings
SEMFIELDfamily_and_sexuality
SEMFIELDthe_human_body
LACKOFSALIENCE.relative.nr.mwe
LACKOFSALIENCE.lack.of.prevale
AFFECT.polaritynegative
AFFECT.polaritypositive
AFFECT.rel.nr.dimin
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 618 degrees of freedom
  (230 observations deleted due to missingness)
Multiple R-squared:  0.3273,    Adjusted R-squared:  0.3186
F-statistic: 37.59 on 8 and 618 DF,  p-value: < 2.2e-16
```

model is only mediocre:
about 32% of variation is explained by model

vs. pilot study:
about 30% more variation explained

Overview

Introduction

Defining concept features

1. semantic field
2. lack of salience
3. vagueness
4. affect

Defining onomasiological heterogeneity

1. diversity
2. geographic fragmentation
3. onomasiological heterogeneity

Methodology & results

Conclusion

Conclusions

replication of pilot study confirms influence of (heterodox) concept features on lexical geographical variation

some differences:

- model explains smaller amount of variation in the data set
- influence of concept vagueness not confirmed

additional predictor:

- semantic field of a concept has significant influence on onomasiological heterogeneity
- not expected effect

→ does semantic field model a hidden variable?

Future directions

- fine-tune predictors:
  - additional (objective) operationalizations of predictors (corpus data)
  - expand to more semantic fields and more dialect areas
- semantic field:
  - why do differences between semantic fields occur?
  - hidden variable?
- response variable:
  - what is the influence of the predictors on different parts of the response variable?

Thank you!

for further information:

karlien.franco@.kuleuven.be

http://wwwling.arts.kuleuven.be/qlvl/karlien

References

Geeraerts, D. and Speelman, D. (2010) Heterodox concept features and onomasiological heterogeneity in dialects. In Geeraerts, D., Kristiansen, G. and Peirsman, Y. (eds.), *Advances in Cognitive Sociolinguistics*: 23-40. Berlin/New York: De Gruyter Mouton.

Keuleers, M., Stevens, M., Mandera,P. & Brysbaert, B. (2015) Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology .*

Speelman, D. and Geeraerts, D. (2008) The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing 2(1–2):* 221–242.

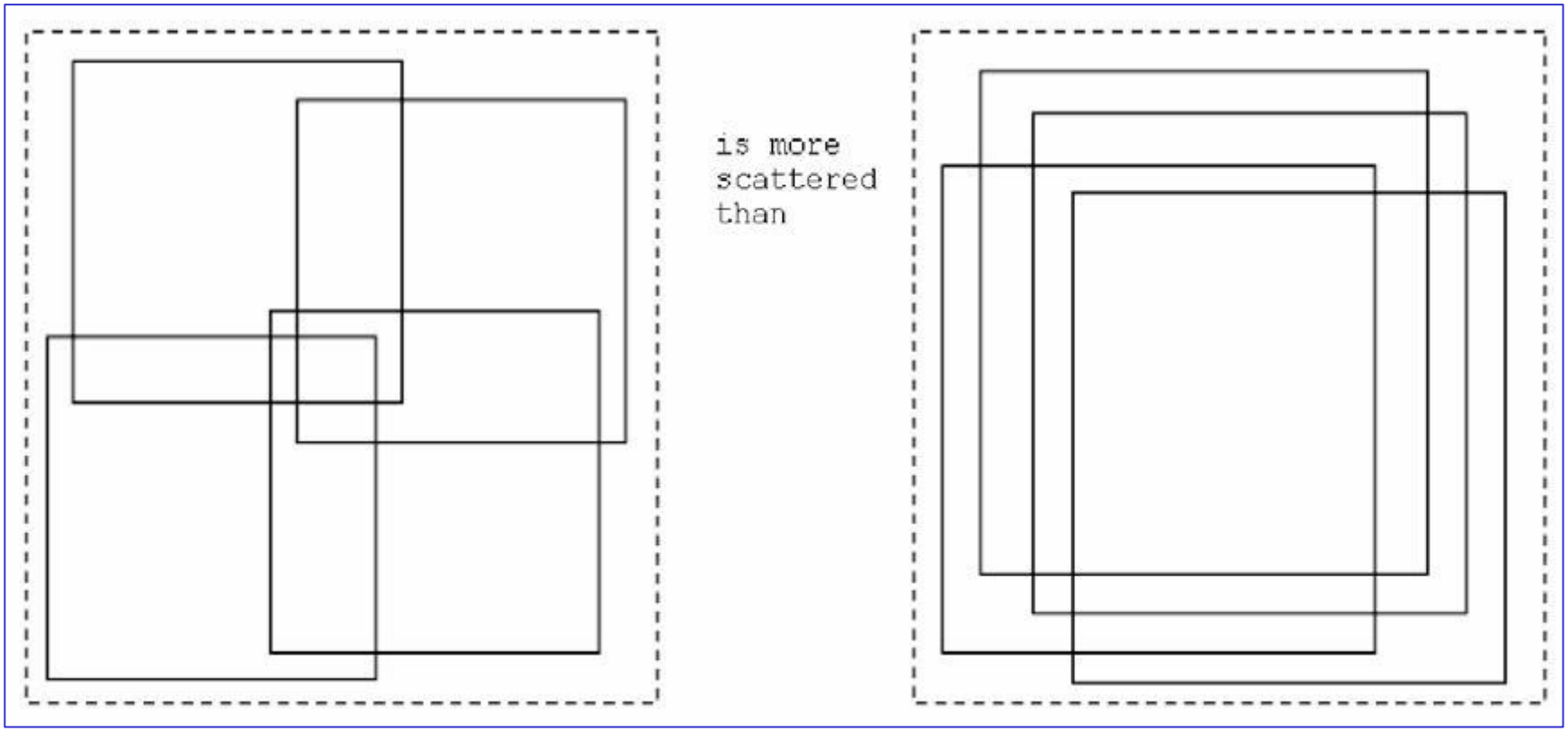
Operationalization onomas. heterogeneity

= log(diversity * fragmentation)

diversity = # types/concept

fragmentation = dispersion/range

Calculation 1/range



Calculation 1/range

= lack of spread (words per concept)

= 1/(weighted)avg. rel. geogr. range of words for a concept

→ avg. rel. geogr. range =

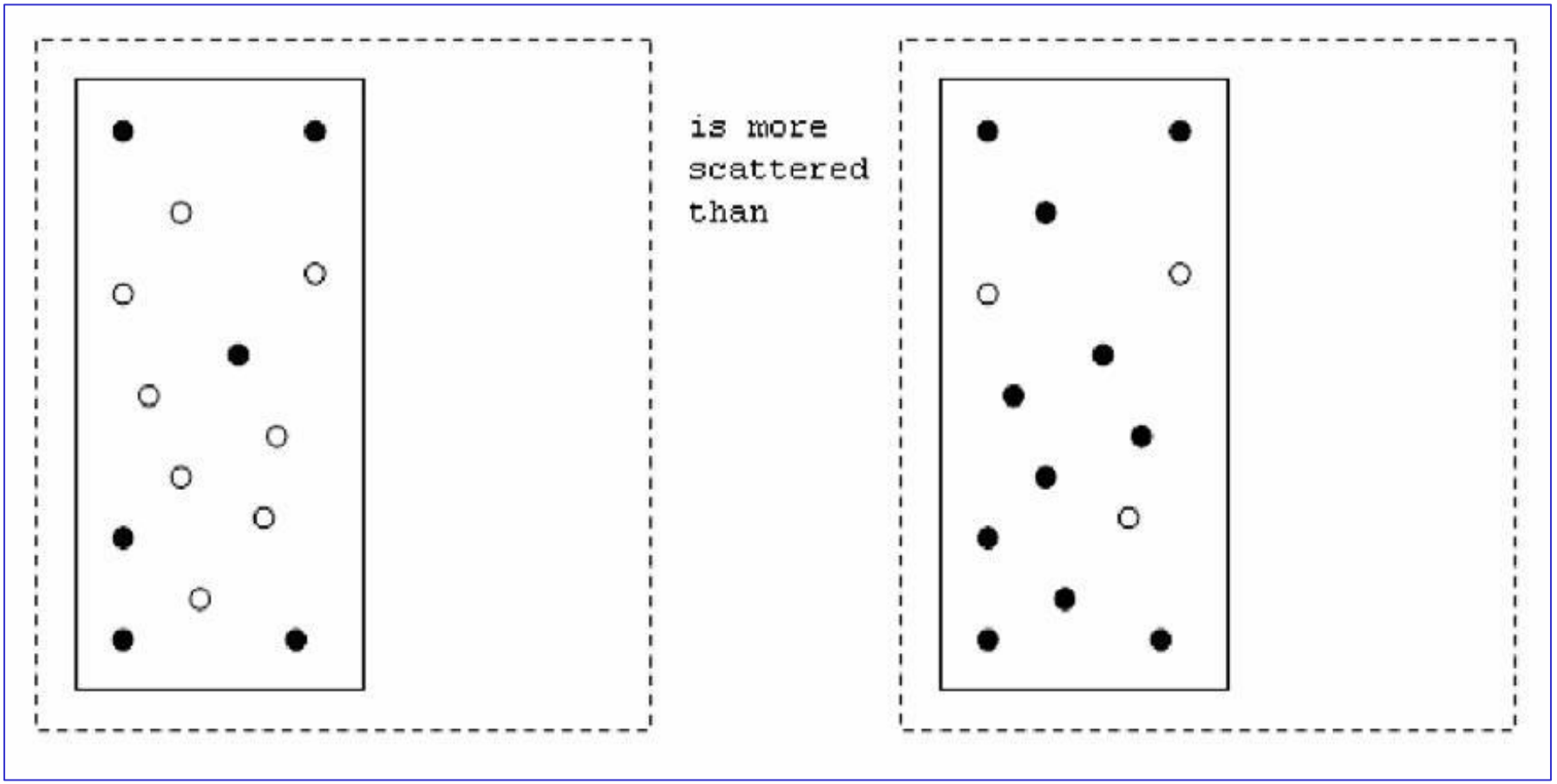
surface of the region where the word is used for the concept

divided by

the surface of the region for which we have records for the concept

weights: words which quantitatively are more important have a higher influence on the measure than words which quantitatively are less important.

Calculation dispersion



Calculation dispersion

= lack of spread (locations in dataset)

calculation

1. calculate word-specific dispersion for each word used to name the concept

→ *numerator*: mean geographical distance from each location where the word is used for the concept to the closest other location where it is also used for the concept

→ *denominator*: mean geogr. dist. from each location where the word is used for the concept to the closest other loc. with records for the concept, but not necessarily for the word

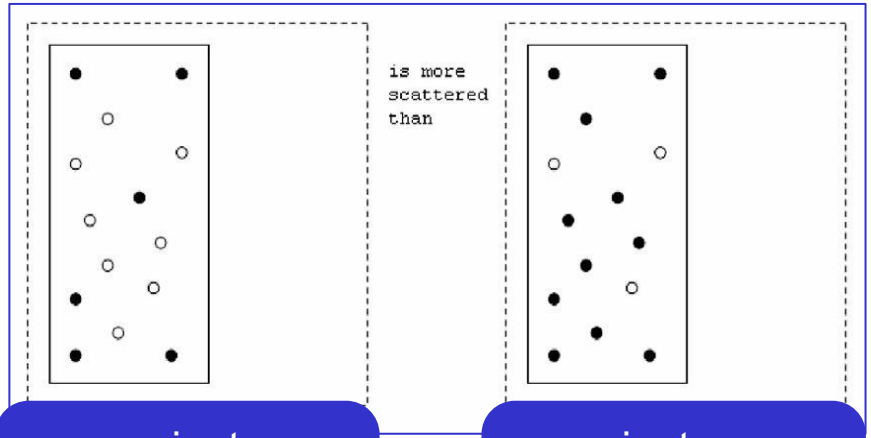
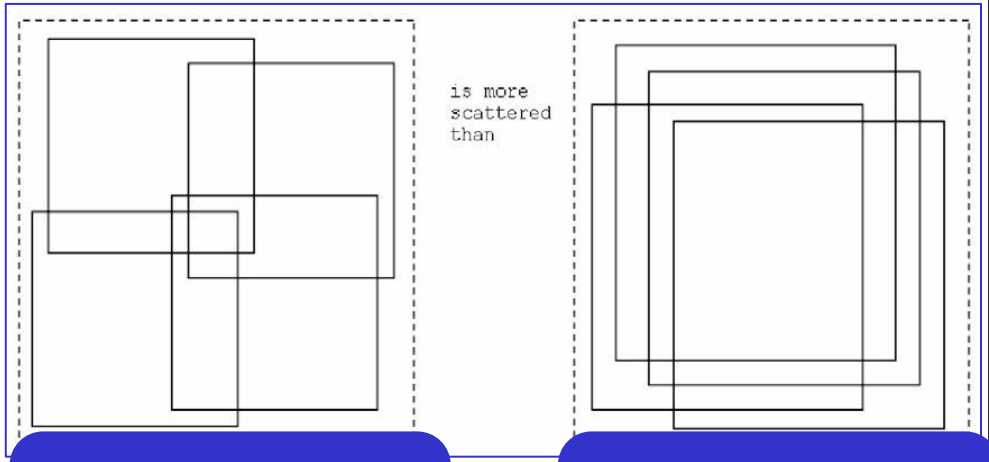
2. global dispersion

= weighted average of word-specific dispersion measures

2. Geographic fragmentation

~ lack of homogeneity in geographical spread of a concept

- consists of two parts:

dispersion	range
	
variants scattered across dialect area	each word type occurs in small geographical area
variants are found in nearby locations	each word type takes up almost entire dialect area